Project Milestone 1

This is to ensure that you are making significant progress towards your project in the class. This is the first milestone with the following deliverables:

1. Data Source, Descriptions and Sample: You should have your dataset finalized by now. Please provide the following information about the dataset that you will be using for the project:
   1. Short description of the dataset (e.g. the dataset describes the sales transactions of a retail store in Europe).
      1. The dataset describes university spending habits in regards to select Division 1 NCAA sports the university participates in. This is supplemented with an additional dataset containing associated win/loss records for the sports teams.
   2. Data source (a link to either a Box or Google Drive folder where you have downloaded the data. Please do not provide the link to the website or the API that provided you with the data).
      1. Win-Loss data is accessible here:

[https://drive.google.com/drive/folders/1UyhaGlulqNVFVZ9ZYE9Inaw](https://drive.google.com/drive/folders/1UyhaGlulqNVFVZ9ZYE9InawhgHQbvEKH?usp=drive_link) hgHQbvEKH?usp=drive\_link

* + 1. Schools and Relevant Financial data is accessible here:

<https://lmu.box.com/s/n3w31o0kg7dst9k0uv98pvnrbakkks9d>

* 1. The total number of tables, columns in each table and the number of rows in each table.
     1. We have created an LMU Build database and are using python to split our data into the columns sets needed for the tables in our database (see ERD). Our progress on this is stored here, along with the associated files:

<https://lmu.box.com/s/e2ql7xn6fyq1lmjto2zefqyxy19ofeii>

There are two python files in this folder, one for each of the survey datasets (finances by school and finances by sport). We advise downloading the entire folder to ensure the code will run without error. Password: BSAN6060NCAA

The original “schools” dataset had 129 columns and, combining all years, we have 215,440 rows. Our relational database will look quite differently, however. Please see the ERD for details.

* 1. A sample (no more than 1000 rows) from the most significant table of the dataset.
     1. This is a sample of the expenses and revenues by sport and institution:

<https://lmu.box.com/s/939klbfn68y5dwwkv1h2qliyqkzdk2ps>

This was derived from the dataset using python.

* 1. Any data quality issue that you have to address (missing data, misspelling, text format for numbers, etc.)

Survey Data

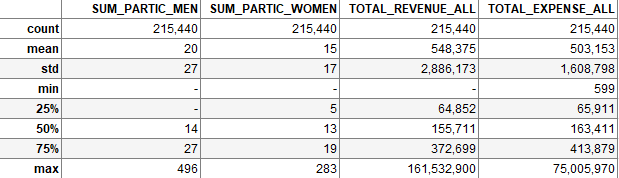
1. The 2022 file has an extra column "OPEID" which had to be stripped to match the column set of the other files
2. The 2007-2009 SAS files have an error. We will have to read these in as .xlsx or .csv files or limit our scope to 2010-2022.
3. There are multiple labels for track and field corresponding to the same codes. Therefore, we will use the code as the primary key in the Sports table to enforce referential integrity. We also notice a label for "all track combined". On investigation, this presented no duplication issues. Schools with this code/label had only this code/label for track.

Win-Loss Data

1. The school names in this dataset were inconsistent with the survey dataset. We created a lookup table to translate the names.
2. The data source lacked data from 2009-2010 across all sports we pulled data for. Women's volleyball and soccer both additionally lacked 2010-2011 data. We will decide as we do our analysis to leave the data as is, scope our project around the missing data, or look for supplemental data sources to fill in the gaps
3. Preliminary descriptive and diagnostic analysis. This should contain:
   1. A summary of some of the items of interest that the project is interested in exploring (mean, median, mode, stdev, etc.)
      1. Our team is interested in comparing win-loss ratios over the past 10-12 years to the expenses incurred by school. We will be looking for patterns in the expenditure amounts by type and school in relation to sports performance.
      2. Our team is also interested in other factors that affect performance, such as number of coaches, number of participants, and location.
      3. We will not be able to study these relationships until our datasets are combined into one schema. However, we have examined the datasets with some preliminary descriptive analytics:
         1. Python was used to study the data:

Win-loss <https://colab.research.google.com/drive/13bfAhJxZCmNO6Eqhe0-8gvdVT5QA85My?usp=drive_link>

Survey Data



* 1. A few simple graphs showing relationships/summary of some of the variables.
     1. Preliminary tableau graphs have been prepared here:

<https://lmu.box.com/s/2eyt7w48vop1r29e5q2ey2mf4fyzgaud>

* 1. Any interesting observation that you think should be explored further.
     1. We observed that the expenses and revenue sharply increased beginning in 2021. We plan to look into the context to understand if the survey definition changed or if there were socioeconomic factors driving this change.
     2. LMU is #10 in the top 10 spenders in their classification (NCAA division I without football). It would be helpful to see how many sports they fund in comparison to others for better context.

1. A very rough list of facts and dimensions that should be included in your dimensional model (star schema). If you are also turning the data into a relational database, then an ERD for that database.

An ERD is attached in brightspace. It is also available here:

<https://lmu.box.com/s/cxefp0riqtdr9tv33og6e2t04aorab4j>

FACTS:

Revenue\_Amount

Expense\_Amount

Win-Loss\_Ratio

Coach\_Count

Participant\_Count

Coach\_Salaries

Student\_Aid

Recruitment\_Expenses

Total\_Operating\_Expenses

DIMENSIONS

Year

Institution

Sport

Location (zip code)

Classification

1. A timeline showing the plan for the project for the rest of the semester along with outcomes that you want to achieve at the end of each period in your plan.

**Data Architecture Design /ELT Development (March 24 - April 3)**

* Design a data architecture that aligns with the identified business problem and data sources.
* Develop our complete star schema.
* Implement Extract, Transform, Load (ETL) processes to integrate and prepare data for analysis.
* Address any remaining data cleansing or transformation requirements identified during the assessment phase.

**Descriptive Analytics and Predictive Analytics (April 3 - April 10)**

* Perform descriptive analytics to gain insights into historical data trends and patterns.
* Explore the feasibility of predictive analytics techniques, if applicable, to forecast future outcomes.

**Visualization Development (April 10 - April 17)**

* Create visualizations using tools like Tableau, Excel, or Python to communicate insights from the data.
* Ensure visualizations are aligned with the business problem and provide actionable insights for decision-making.

**Final Report Preparation (April 17 - April 24)**

* Compile findings, analysis, and visualizations into a comprehensive report or slide deck for the final presentation.
* Include slide notes and annotations to provide context and explanation for each component.

*I would also recommend that you start thinking about the final dashboard that you want to build as well as the overall dataflow for loading the data sources that would feed the dashboard. You can use Whimsical for creating your flow diagram. It is a free tool and allows collaborative drawing.*